Carbon—The First Frontier of Information Processing

Apoorva Patel
CTS and SERC, Indian Institute of Science, Bangalore-560012
E-mail: adpatel@cts.iisc.ernet.in
(5 March 2001)

Information is often encoded as an aperiodic chain of building blocks. Modern digital computers use bits as the building blocks, but in general the choice of building blocks depends on the nature of the information to be encoded. What are the optimal building blocks to encode structural information? This can be analysed by replacing the operations of addition and multiplication of conventional arithmetic by translation and rotation. It is argued that at the molecular level, the best component for encoding structural information is carbon. Living organisms discovered this billions of years ago, and used carbon as the back-bone for constructing proteins which function according to their structure. Structural analysis of polypeptide chains shows that 20 building blocks are necessary to fold them into arbitrary shapes. Properties of amino acids suggest that the present genetic code was preceded by a more primitive one, coding for 10 amino acids using two nucleotide bases.

PACS: 03.67.Lx, 87.15.By

I. STRUCTURAL INFORMATION

It is a characteristic of living organisms to acquire information, interpret it and pass it on, often using it and refining it along the way. This information can be in various forms or languages. It can be genetic information passed on from the parent to the offspring, sensory information conveyed by the sense organ to the brain, linguistic information communicated by one being to another, or numerical data entered in a computer for later use. It is advantageous to process the information efficiently, and not in any haphazard manner. In case of living organisms, Darwinian selection during evolution can be considered the driving force for such optimisation. In general, information processing is optimised following two guidelines: minimisation of physical resources (time as well as space), and minimisation of errors.

A striking feature of all the forms of information listed above is that the messages are represented as aperiodic chains of discrete building blocks. Such a representation, called digitisation of the message, is commonplace due to its many advantages. Discretisation makes it possible to correct errors arising from local disturbances, and so it is desirable even when the underlying physical variables are continuous (e.g. voltages and currents in computers). It is also easier to handle several variables each spanning a small range than a single variable covering a large range. Any desired message can then be constructed by putting together as many as necessary of the smaller range variables, while the instruction set required to manipulate each variable is substantially simplified. This simplification means that only a limited number of processes have to be physically implemented leading to high speed computation. An important question, therefore, is to figure out the best way of digitising a message, i.e. what should be selected as the building blocks of the aperiodic chain.

The information contained in a message depends on the values and locations of the building blocks. Given a set of building blocks, Shannon quantified the information contained in a message as its entropy, i.e. a measure of the number of possible forms the message could have taken. This measure tells us that the information content of a message can be increased by eliminating correlations from it and making it more random. It also tells us that local errors in a message can be corrected by building long range correlations into it. But it does not tell us what building blocks are appropriate for a particular message. The choice of building blocks depends on the type of the information and not on the amount of information.

Information can be translated from one language into another by replacing one set of building blocks used to encode the information by another, e.g. textual information is stored in the computer in a binary form using the ascii code. Nonetheless, physical principles are involved in selecting different building blocks for different information processing tasks. For example, our electronic computers compute using electrical signals but store the results on the disk using magnetic signals; the former realisation is suitable for quick processing while the latter is suitable for long term storage. In selection of building blocks with appropriate properties, the foremost practical criterion is that it should be easy to distinguish one building block from another. We use decimal system of numbers because we learnt to count with our fingers. The number of syllables in our languages are determined by the number of distinct sounds our vocal chords can make. Computers and nervous systems use binary code because off/on states can be quickly decided with electrical signals. Genetic information is encoded using four nucleotide bases, perhaps because the quantum assembly algorithm is the optimal choice for replication occurring at the molecular scale [1].

Numerical representation of information is one-dimensional and uses building blocks with an ordering amongst them (e.g. one is greater than zero). But these

features may not be present in other types of information. For example, ordering is not required for letters of an alphabet, and it may be possible to represent structural information by higher dimensional building blocks. The information in the genes for the synthesis of proteins is a clear-cut example of structural information. The shape and size of a protein determines its role in biochemical processes, much more so than its chemical content. The 3-dimensional structural information of proteins is encoded as a one-dimensional chain of building blocks—the amino acids—with the interaction amongst the building blocks determining how the chain would bend and fold. It is thus natural to ask: what is the best way of encoding structural information? This is the question addressed in this work.

Any structural transformation of a rigid body can be described in terms of two basic operations, translations and rotations. (For non-rigid bodies deformations are possible structural transformations, but deformable objects are not very useful for encoding structural information and I leave them aside.) The set of all rigid body translations and rotations form the well-known Galilean group, which has been studied in detail by physicists. To construct the building blocks, we have to discretise this continuous group and yet maintain its features required to encode structural information.

We can compare translations and rotations to the fundamental operations of arithmetic—addition and multiplication. While addition is nothing but translation along the real line, multiplication is quite different from rotation. Rotations in our 3-dimensional space are not commutative [2], and that is of crucial importance in representing structural information. The building blocks of numerical information are elements of Z_n, the group of integers modulo n, and the cyclic nature of this group represents the order amongst the building blocks. The building blocks of structural information need to have characteristics of rigid bodies, i.e. specific size and orientation in 3-dimensional space. To find them we have to look for a finite non-commutative group. In addition, to address the question of protein structure, we should look for transformations that take place at the atomic scale.

Translations are easily discretised, as uniformly spaced units along a polymer chain. The atomic structure of matter provides a natural unit for translation—the physical size of the building blocks. Indeed, the amino acids making up proteins differ from each other in terms of their side chemical groups, while their components along the chain are identical. Any translation can be built up from the elementary operations of addition of a building block, deletion of a building block and exchange of two adjacent building blocks.

Rotations are more complicated to discretise. A reasonable criterion is to demand, on the basis of symmetry, that the allowed states be all equivalent and equidistant from each other. The largest set of such states can then provide an approximate basis for the rotation group, and the following properties are quickly discovered:

• In our 3-dimensional world, the largest number of equivalent and equidistant states is four. They are the corners of a regular tetrahedron. Any one of the states can be rotated to any other with equal ease.

• A tetrahedron is the smallest polyhedron. It is the simplest structure that can implement non-commutative features of 3-dimensional rotations.

• To be able to specify the 3-dimensional orientation unambiguously, the building blocks should have the capability to include a chiral center.

• If quantum dynamics is involved, then the states should also be mutually orthogonal, so that they form a basis for the Hilbert space. The tetrahedral quantum states are mutually orthogonal; they can be obtained from the lowest two spherical harmonics, $l = 0$ and $l = 1$. ($l = 0$ alone is isotropic; inclusion of the next l value gives the minimal basis set for specifying orientations.) Using sp^3-hybridisation of atomic orbitals, these states can be denoted as:

$$\begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & -1/2 & -1/2 \\ 1/2 & -1/2 & 1/2 & -1/2 \\ 1/2 & -1/2 & -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} s \\ p_x \\ p_y \\ p_z \end{pmatrix} \quad (1)$$

The high symmetry of this unitary transformation (all elements equal, only signs differ) is related to the equivalence of the four states.

• Four is also the largest number of states which can be uniquely identified by a single yes/no question in a quantum search algorithm [3].

II. TETRAHEDRAL GEOMETRY

The outstanding example of an element with such states is carbon. Moreover,

◦ Carbon has the capability to form aperiodic chains, where different side chemical groups hang on to a backbone. This capability is a must for encoding information. Silicon also possesses the same tetrahedral states, and is much more abundant, but it preferentially forms periodic chains (i.e. regular crystals).

◦ If the logic above is repeated in the case of 2-dimensional rotations, it leads to three equivalent states located at the corners of an equilateral triangle. Carbon has the capability to form these states as well, by sp^2-hybridisation of its atomic orbitals.

◦ Carbon is the most important structural element forming the back-bone of biomolecules. Darwinian selection in evolution can be expected to have picked the best building blocks out of the available resources.

With all these pieces fitting together, let us look at the tetrahedral group in some detail. The tetrahedral group is isomorphic to the permutation group of four objects. It has 24 elements, which can be factored into a group of 12 proper rotations (or even permutations) and reflection (or parity). The 24 element and 12 element groups are denoted as T_d and T respectively.

A regular tetrahedron can be formed by joining alternate corners of a cube. The centres of the tetrahedron and cube then coincide, and this embedding is convenient for structural analysis of a 3-dimensional chain with tetrahedral angles. The 12 proper rotations are decomposed into the identity operation, rotations around 3-fold axes and rotations around 2-fold axes. There are four 3-fold axes, joining the centre of the tetrahedron with a vertex; $+120^\circ$ and -120° rotations around these axes belong to different equivalence classes. There are three 2-fold axes, passing through the center of the tetrahedron and midpoints of its non-intersecting edges (equivalently passing through the centres of opposite faces of the embedding cube).

For a carbon atom located at the centre of the tetrahedron, rotations around 3-fold axes correspond to rotations around its bonds. These single bonds are easy to rotate and give rise to different conformations of organic molecules. In a polypeptide chain, the orientations that can be achieved by rotations around the bonds of the C_α atoms are described by the Ramachandran map. As shown in Fig.1, the rotation angles are not uniformly populated, but prefer to be in several discrete locations. As the stars in the plot show, discretising the angles in steps of 120° is not a bad starting point.

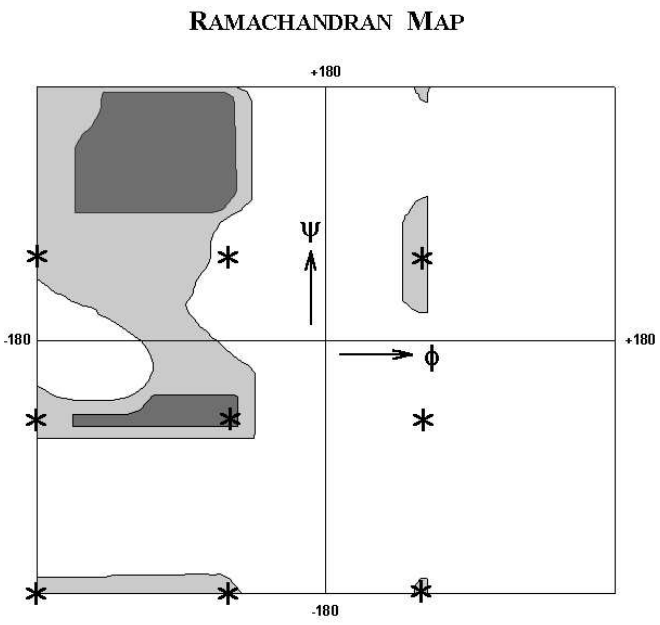

FIG. 1. The Ramachandran map displaying the distribution of rotation angles for the C_α bonds in polypeptide chains (courtesy C. Ramakrishnan) [4]. The angles ϕ and ψ are periodic. In the approximation that embeds the polypeptide chain on a diamond lattice in the "trans" configuration, only nine discrete possibilities exist for the rotation angles. These are marked by stars on the same plot; they are uniformly separated by 120° steps.

The 2-fold rotation axes bisect the bond angles. If a double bond is viewed as a deformation in which two tetrahedral bonds are merged together, then the double bond lies along the 2-fold rotation axis. 180° rotation about this axis corresponds to a transition between "trans" and "cis" forms. Most of the peptide bonds have the "trans" configuration. But occasional transitions to the "cis" form do occur, and they are important for introducing sharp bends in the chain.

The parity transformation flips chirality of a structure, which is of special significance for many biological molecules. Chirality flip is an allowed quantum transformation, e.g. the NH_3 molecule flips back and forth between configurations where the nitrogen atom is above and below the plane of three hydrogen atoms. But chirality flip becomes more difficult as the molecular size increases, and all the amino acids used as building blocks of proteins are known to be L-type (except for achiral glycine). Thus reflections are more difficult to implement than proper rotations, and can be ignored as far as the structural analysis of proteins is concerned.

III. PACKING 3-DIMENSIONAL INFORMATION

Multi-dimensional structural information can be encoded in several different ways. The complete information can be expressed directly, as in holograms (3-dim) and movie projections (2-dim). Or it can be arranged as an ordered set of lower dimensional segments, as in CT-scan (stack of parallel planes covering a 3-dim object) and television monitors (set of lines covering a 2-dim picture). The choice depends on whether the physical means that convey the information are extended or local. When mechanisms exist to look at the whole object in one go (e.g. with a wide beam of light), the complete information can be addressed directly. When only one part of the object can be considered at a time (e.g. with a narrow beam of electrons), it is more convenient to arrange the information as a sequence of small segments. When the building blocks themselves have to convey the information, the latter format is the obvious choice; multi-dimensional arrays are stored as a folded sequence in computers and proteins are assembled as folded polypeptide chains.

It is possible to assemble arbitrary structures by repetitive arrangement of a single and small enough building block. For example, a crystal can be carved into the desired shape, and it is sufficient to describe the details of the surface (and not the contents of the full volume) for that purpose. A crystal can also form rapidly, since it can grow from a seed in all directions. But the preferred shape of the crystal remains that of the building block. To assemble arbitrary shapes using a crystalline arrangement, another agency is needed to tell the crystal surface to stop growing after it has reached the desired position as well as to put the reactive chemical groups at

specific locations; the building blocks themselves cannot carry those instructions. Thus crystal growth is convenient for making regular patterns, but it is not a good choice for assembling irregular shapes. The highly nontrivial task of specifying the surface shape can be more easily accomplished by an aperiodic folded chain of building blocks. The building blocks themselves carry preferences for specific orientations at each step; although such a chain grows slowly, it does not need help from another agency to achieve its desired shape. Proteins need all their grooves and cavities (i.e. defects in the structure) for their function, and how they fold is decided by the amino acid sequences. Another physical reason why proteins have to be polypeptide chains that can unfold and fold again is described in the next section.

Even after picking a folded chain structure, more specifications are needed to find the desired building blocks. The chain can be uniformly flexible like a piece of string, or it can be made of stiff segments alternating with flexible joints like a chain of metal rings. If all the segments of a chain are flexible, then it has to be fully tied from all directions to be held in place. Otherwise the structure can crumple and collapse. Carbon forms many structures with fully saturated bonds, but a completely tied 3-dimensional form requires rather precise folding and cannot accommodate aperiodic building blocks easily. For example, diamond is the hardest material, but it is a periodic structure and cannot be unfolded and folded again easily. The polyethylene back-bone (i.e. $(-CH_2-)_n$) can accommodate aperiodic side groups, but it is too flexible. Rigidity of the back-bone can be increased by converting some of its single bonds into double bonds that cannot rotate. Polypeptide chains indeed are of that type; it has peptide bonds that cannot rotate, and the increased stiffness helps in maintaining the shape of the protein.

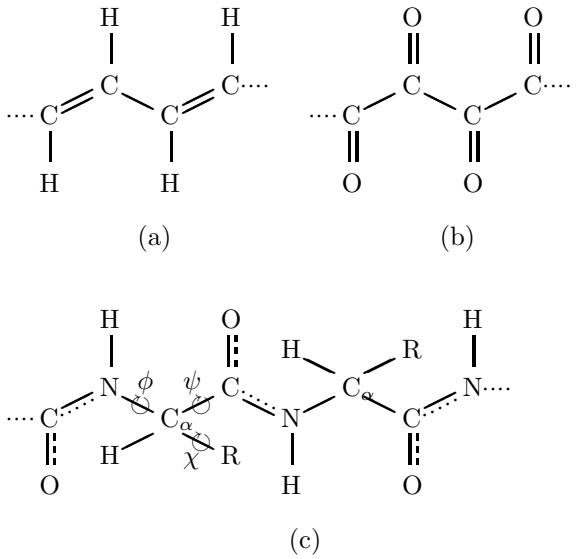

FIG. 2. Different possibilities for polymer chains with carbon back-bone: (a) $(CH)_n$, (b) $(CO)_n$, (c) polypeptide chain.

Fig.2a shows a back-bone with alternating single and double bonds. This is the structure of polyacetylene, and an aperiodic chain can be constructed by replacing the side $-H$ by other chemical groups (e.g. $-CH_3$). The trouble with this structure is that the $\pi-$electrons involved in double bonds prefer to lower their energy by spilling over into neighbouring bonds. This resonance phenomenon gives a double bond character to all the bonds (the actual bond properties are somewhere in between a single and a double bond), and makes the whole back-bone planar. A planar back-bone is no good for constructing 3-dimensional structures. The double bonds can be shifted to the side groups to reduce the spill over of $\pi-$ electrons, as illustrated in Fig.2b, but the resultant structure is still planar. The next possibility for the back-bone configuration, which allows stiff segments with flexible joints, is to alternate one double bond with two single bonds. This is the structure of polypeptide chains, as shown in Fig.2c. The stiff $C-N$ peptide bond is created by $\pi-$electrons spilling over from the $C=O$ double bond, and rotatable single bonds permit construction of 3-dimensional structures.

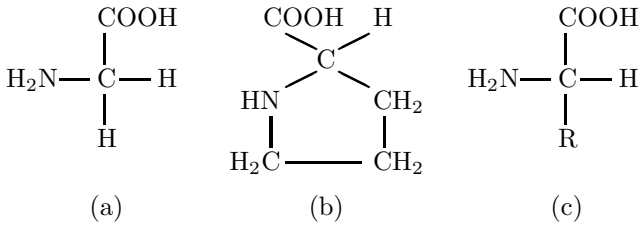

FIG. 3. Amino acid configurations: (a) glycine, (b) proline, (c) all the rest.

IV. PROTEINS

Having analysed the merits of a polypeptide back-bone structure, let us review some of the important properties of proteins [5,6]:

• The sequence of amino acids encodes the structural information of a protein. Fig.3 shows the structure of individual amino acids. Proteins may include other important components, e.g iron in haemoglobin, but role of these other components is essentially chemical and not structural.

• The sequence of amino acids is obtained by translation from the sequence of nucleotide bases in DNA. This translation is necessary because the two languages serve different purpose, and their purpose decides the physical components for their realisations. According to the cell's need, proteins are synthesised, transported to appropriate location to participate in biochemical reactions, and degraded at the end. The shape of the protein plays a critical part in its reactivity. The double helical structure of DNA, with nucleotide bases hidden inside, protects the information until it is required. DNA replication is also much less error-prone than protein synthesis.

• The physical separation between consecutive amino acids in polypeptide chains and consecutive nucleotide bases in DNA is about the same, approximately $3.5\mathring{A}$. It is not the stereochemistry, therefore, which is responsible for three nucleotide bases being mapped to one amino acid in the genetic code. The non-overlapping triplet code is likely to have arisen from the need to have a sufficient number of amino acids as the required building blocks for the 3-dimensional protein structure. Living organisms had to then set up the complex machinery, involving tRNA as adapters connecting nucleotide bases and amino acids, to carry out the task of translation.

• Correct translation is ensured by the bilingual aminoacyl-tRNA synthetases that attach amino acids to tRNA molecules with appropriate anticodons. There may be several anticodons which map to a particular amino acid, but there is only one aminoacyl-tRNA synthetase per amino acid which carries out the many-to-one mapping. Once the tRNA molecules are properly charged with amino acids, the ribosomes match the anticodons of tRNA with codons of mRNA and construct the polypeptide chain.

• Atoms in proteins are quite densely packed. In terms of the van der Waals atomic size, packing fraction for proteins is in the range $0.70 - 0.78$, compared to 0.74 for closest packing of identical spheres. The packing density of a diamond lattice is only 0.34, and the side groups of a polypeptide chain folded along a diamond lattice fill up the empty spaces. Even then the packing density is high along the chain, while the amino acid side groups are somewhat loosely packed.

• Proteins also have to cross membranes and cell walls after their synthesis, since they often have to carry out their task at another location. Membranes and cell walls cannot afford to have big holes, and that provides an important reason why proteins are folded chains. During translocation, proteins unfold to their chain form, cross the barrier and then fold again into their native 3-dimensional form. The ability of proteins to unfold and fold again requires that the bonds along the chain be much stronger than the side bonds that hold the folds in the chain.

• Certain domains of proteins start folding as soon as they are synthesised, indicating that at least some of the folding rules are local. The 3-dimensional protein structure is essentially dictated by the sequence of amino acids. Globular proteins can be melted by heat, and they regain their native form upon cooling.

• Carbon and nitrogen atoms, joined by strong covalent bonds, form the back-bone of the polypeptide chain. In this chain rigid peptide bonds alternate with rotatable bonds of C_α atoms. The $C - N$ peptide bonds have a double bond character; the nitrogen atom carries a positive charge making its electronic behaviour similar to the tetravalent carbon atom. Different amino acids are distinguished from each other by their $R-$groups, which are side chains attached to the C_α atoms.

• Amino acid $R-$groups are of various types: polar, non-polar, aromatic, positively and negatively charged. The interactions of these $R-$groups with each other and with the ambient water molecules fix the orientations of the rotatable C_α bonds. These interactions are weak and easily influenced by the pH of the ambient liquid and the temperature.

• The folding process occurs in stages. Local domains fold first, essentially due to weak bonds (Hydrogen and van der Waals). This process is dominated by local transformations, i.e. proper rotations of bonds of C_α atoms, and forms well-known structures such as $\alpha-$helices and $\beta-$sheets. In the next stage, already folded domains get linked by long-distance transformations, e.g. disulfide bonds. In the final stage, various separately assembled structures, polypeptide chains and chemical groups, join together.

• Regular structures like $\alpha-$helices and $\beta-$sheets are largely determined by the properties of the polypeptide back-bone, with a lot of freedom in choice of amino acid $R-$groups. It is the irregular twists and turns of the chain which critically depend on the interactions of the amino acid $R-$groups.

• In reality, proteins fold rather rapidly. The folded chain is a self-avoiding walk in 3-dimensional space. Such a walk can get stuck for topological reasons, or it may need global criteria to complete its task (travelling salesman type of problems are NP-hard with just local rules). An easy escape is to complete the task with multiple walks, i.e. start a new walk when the previous one gets stuck. Indeed many proteins are made of not a single polypeptide chain, but several polypeptide chains entangled together. Large protein structures are often made of polypeptide units arranged in regular patterns (e.g. fibroin, keratin, collagen, virus coats etc.).

• Fig.2c shows the rotation angles around single bonds in a polypeptide chain. Because of steric conflict between various atoms, not all the values of angles occur in a polypeptide chain. The Ramachandran map ($\phi-\psi$ angular distribution) in Fig.1 displays the orientations available to the amino acids. Experimental data for polypeptide chains follow the constraints of the Ramachandran map quite well; in fact, the map is often used as a filter for the experimental data. The side chain angles (χ) also have preferred orientations which are separated by $120°$.

• Structural roles played by some of the amino acids are well-known. Glycine with no side chain and no chiral centre is the most flexible. Proline with its rigid ring and trans-cis transformation plays an important role in forming sharp bends. Cysteine connects far separated regions of the polypeptide chain by strong disulfide bonds, helping the folded chain retain its shape.

V. ELEMENTARY BUILDING BLOCKS

We now look at the 3-dimensional geometry of a polypeptide chain, but with the simplifying assumptions

that all the links in the chain are of equal length and all the tetrahedral angles are of equal value ($2\tan^{-1}(\sqrt{2}) \approx 109.5°$). With these assumptions, the folded chain lies on a diamond lattice. Although the real peptide bond is planar with angles of 120°, it can be fitted reasonably well on the diamond lattice in the "trans" configuration (i.e. two outgoing bonds parallel). The rare "cis" configuration, obtained from the "trans" configuration by a 2-fold rotation around the peptide bond, takes the chain out of the diamond lattice. Let us first keep that aside, and consider the chain in "trans" configuration only. In a real polypeptide chain, variations from equal bond lengths and equal angles are within ±10%, and I will analyse the above described simplified version using the conventional polypeptide chain nomenclature.

The diamond lattice is a face-centred cubic lattice with a two-point basis. Let this basis of lattice points be $(0,0,0)$ and $(1/4,1/4,1/4)$ in units of the unit cell. Then the bond directions of the diamond lattice are (these are the last three columns of the matrix in Eq.(1)):

$$\begin{aligned} e_1 &= (+1/4, +1/4, +1/4) \\ e_2 &= (+1/4, -1/4, -1/4) \\ e_3 &= (-1/4, +1/4, -1/4) \\ e_4 &= (-1/4, -1/4, +1/4) \end{aligned} \qquad (2)$$

These directions refer to the lattice point at the origin, and thereafter the bond directions at adjacent points are opposite in sign.

We can enumerate all possible configurations of the polypeptide chain, by specifying for every peptide bond the location of the next peptide bond in the chain. Let the reference peptide bond $(C-N)$ be along e_1 from the origin. The chain prior to this peptide bond is already synthesised, so without loss of generality let the location of the C_α preceding the reference peptide bond be e_2. In the "trans" configuration, the $N-C_\alpha$ and the $C_\alpha - C$ bonds are parallel, so the location of the C_α following the reference peptide bond is fixed as $e_1 - e_2$. (The sequence $C_\alpha - C - N - C_\alpha$ fixes the plane of the peptide bond.)

There are three possible locations for the next C: $e_1 - e_2 + e_1$, $e_1 - e_2 + e_3$ and $e_1 - e_2 + e_4$. From each of these three locations, the next peptide bond can proceed along three possible directions, excluding the already occupied $C_\alpha - C$ direction. The $C_\alpha - C$ direction and the next peptide bond direction fix the plane of the next peptide bond. Thus on a diamond lattice, given a peptide bond plane, there are 9 possible positions for the next peptide bond plane.

For each of the polypeptide back-bone configuration described above, there are two remaining directions for other groups to attach to the C_α atom. One direction is attached to the R-group of the amino acid, while the other to a hydrogen atom. There are two arrangements possible, and they correspond to opposite chirality. Detailed model-building studies have shown that all the R-groups in a polypeptide chain must be of the same stereoisomer for the stability of regular secondary structures (e.g. $\alpha-$helices and $\beta-$sheets). All the amino acids naturally occurring in proteins are L-type. Altogether, therefore, there remain 9 possible ways of adding a new amino acid to an existing polypeptide chain [7].

The Ramachandran map shown in Fig.1 is constructed using the bond lengths and angles in an actual polypeptide chain. The nine points corresponding to the "trans" configuration discrete chain are marked as stars on the same plot. It is easily seen that the discrete approximation is not too far off reality, even though it cannot describe all the details of the $\phi - \psi$ angular distribution. Actually, the plot in Fig.1 does not include glycine and proline; their structural preferences are somewhat different. The region around $(\phi = 60°, \psi = -60°)$ is not occupied in the Ramachandran map because of steric conflict between the side chain R-group and the atoms in the polypeptide back-bone. Glycine with no side chain does not have this conflict and can occupy this region—its Ramachandran map has inversion symmetry. In case of proline, the rigid imino ring does not allow the $N - C_\alpha$ bond to rotate, and ϕ is constrained to be around $-60°$. In case of a real polypeptide chain, embedding it on the diamond lattice will distort its shape; the extent of distortion will then be a measure of usefulness of the discretised description [8].

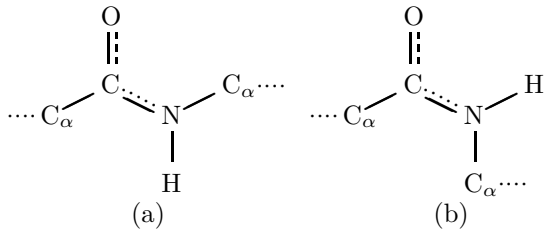

FIG. 4. Peptide bond configurations: (a) trans, (b) cis.

Now we can look at the "cis" configuration of the peptide bond. It is obtained from the "trans" configuration by rotating the $N-C_\alpha$ bond by 180° around the peptide bond axis (see Fig.4). With the peptide bond along e_1 and the preceding $C_\alpha - C$ bond along $-e_2$, the "cis" configuration $N-C_\alpha$ bond is along $\frac{2}{3}e_1+e_2$. This orientation does not fit in the face-centred cubic diamond lattice, but it can be fitted in the hexagonal diamond lattice [9], with the hexagonal symmetry axis along e_1. It is well-known that the 3-dimensional closest packing of spheres can be viewed as a stack of 2-dimensional layers. There are three possible positions for the layers, and each layer has to be displaced relative to the ones on its either side. There are, therefore, two distinct ways to add a new layer onto an existing stack. The face-centred cubic lattice corresponds to the layer sequence $\ldots ABCABCABC \ldots$, the hexagonal lattice corresponds to the sequence $\ldots ABABAB \ldots$, and random sequences are also possible. An insertion of a "cis" peptide bond in an otherwise "trans" peptide chain corresponds to a flip in the layer sequence of the type $\ldots ABCABCBACBA \ldots$. This flip has no effect on the 9 possibilities for the subsequent rotation angles ϕ and ψ, and further elongation of the polypeptide chain. Thus

we can count the trans-cis transformation as one more elementary structural operation.

VI. PUTTING THINGS TOGETHER

There is no clear association of any amino acid with the 9 discrete points in the Ramachandran map; even the most rigid proline occurs in different orientations. Indeed just the physical structure of a particular amino acid does not decide its configuration in the polypeptide chain; rather the overall interactions of its R-group with those that preceded it and those that follow it fix the configuration. The R-group properties of amino acids have been studied in detail: polar and non-polar, positive and negative charge, straight chains and rings, short and long chains, and so on. Still which sequence of amino acids will lead to which conformation of the chain is an exercise in coding that has not been solved yet. Nevertheless, if the coding is efficient than 9 amino acids (may be 10 to include trans-cis transformation) should suffice to form a polypeptide chain of arbitrary configuration.

Amino acid	R-group property	Mol. wt.	Class
Gly (Glycine)	Non-polar	75	II
Ala (Alanine)	Non-polar	89	II
Pro (Proline)	Non-polar	115	II
Val (Valine)	Non-polar	117	I
Leu (Leucine)	Non-polar	131	I
Ile (Isoleucine)	Non-polar	131	I
Ser (Serine)	Polar	105	II
Thr (Threonine)	Polar	119	II
Asp (Asparagine)	Polar	132	II
Cys (Cysteine)	Polar	121	I
Met (Methionine)	Polar	149	I
Gln (Glutamine)	Polar	146	I
Asp (Aspartate)	Negative charge	133	II
Glu (Glutamate)	Negative charge	147	I
Lys (Lysine)	Positive charge	146	II
Arg (Arginine)	Positive charge	174	I
His (Histidine)	Ring/Aromatic	155	II
Phe (Phenylalanine)	Ring/Aromatic	165	II
Tyr (Tyrosine)	Ring/Aromatic	181	I
Trp (Tryptophan)	Ring/Aromatic	204	I

TABLE I. Properties of the 20 amino acids naturally occurring in proteins depend on their side chain R-groups. Larger molecular weights indicate longer side chains. The amino acids have been divided into two classes of 10 each, depending on the properties of aminoacyl-tRNA synthetases that bind the amino acids to tRNA. It is obvious that these same classes divide amino acids with each R-group property equally, the longer side chains correspond to class I and the shorter ones correspond to class II. Some specific properties not explicit in the table are: aspargine is a shorter side chain version of glutamine, histidine has a positively charged R-group but it is close to being neutral, and both the sulphur containing amino acids (cysteine and methionine) belong to class I.

At this stage, it is instructive to observe that the 20 amino acids are divided into two classes of 10 each, according to the properties of their aminoacyl-tRNA synthetases [10–12]. The two classes of synthetases differ from each other in terms of their active sites and how they attach amino acids to the tRNA molecules. The lack of any apparent relationship between the two classes of synthetases has led to the conjecture that the two classes evolved independently, and early form of life could have existed with proteins made up of only 10 amino acids of one type or the other. A closer inspection of the R-group properties of amino acids in the two classes reveals that each property (polar, non-polar, ring/aromatic, positive and negative charge) is equally divided amongst the two classes, as shown in Table 1. Not only that, but the heavier amino acids with each property belong to class I, while the lighter ones belong to class II. This division of amino acids according to the length of their side chains has unambiguous structural significance. The diamond lattice structure is quite loosely packed with many cavities of different sizes. The use of long side chains to fill up big cavities and short side chains to fill up small ones can produce a dense compact structure. Thus we arrive at a structural explanation for the 20 amino acids as building blocks of proteins, 10 for folding the polypeptide backbone and a factor of 2 for the length of the R-group.

A look at the optimal solutions of the quantum search algorithm [3] brings out another interesting feature. Identification of the nucleotide base-pairing with a binary quantum query provided two significant results for genetic information processing [13]: the largest number of items that can be distinguished by one quantum query is 4, and with three quantum queries it is 20.2. The same algorithm also predicts that the largest number of items that can be distinguished by two quantum queries is 10.5 [14]. Therefore two nucleotide bases are optimal for distinguishing 10 amino acids.

The experimentally observed wobble rules are consistent with the idea that an earlier genetic code used only two nucleotide bases of every codon and synthesised a smaller number of amino acids [15]. Another feature supporting this idea is that in the present genetic code similar codons code for amino acids with similar R-group properties. It is therefore possible that the third codon entered the present genetic code as a class label (classical and not quantum), when two independent codes corresponding to long and short R-groups merged together during course of evolution [16].

Combining all these arguments, we can now construct a possible scenario of how the present genetic code arose from a more primitive one (it is remarkably similar to what Francis Crick proposed many years ago [17]):
(1) The primitive code was a triplet one due to some unidentified reasons. The first two letters coded for 10 amino acids, while the third letter was a non-coding separation mark. The individual genes were separate and not joined together, and so START and STOP signals were not needed.

(2) This primitive code synthesised the simpler class II amino acids. The information about how the polypeptide chain twists and turns at each step was incorporated in this code. The short side chains of class II amino acids, however, could not completely fill all the cavities in the 3-dimensional protein structure.
(3) The longer class I amino acids replaced the short ones of similar property at a later stage, wherever big cavities existed. This filling up of cavities increased the structural stability of proteins.
(4) The third letter was put into use as a double-valued classical label for the amino acid class. That allowed coding for 20 amino acids.
(5) Further optimisation of the code occurred with some juggling around of codons, since 20 amino acids can be coded either by one classical and two quantum queries or by three quantum queries. Also, many genes joined together and START and STOP signals were inserted.
(6) Similar codons for similar amino acids and the wobble rules are relics of the doubling of the genetic code, indicative of the past but no longer perfectly realised.
The most important criterion in this scenario is that continuity has to be maintained in evolution—a drastic change will not permit the organism to survive. In the absence of any knowledge of the doublet code, it is not possible to pin-point a particular scenario, and other scenarios can also be imagined [18].

Further progress along this direction requires solutions of two puzzles. First, as already pointed out above, we need to identify which amino acid subsequence corresponds to which structural building block. There is no clear criterion regarding how long the amino acid subsequence should be before it assumes a definite shape; may be interactions of an amino acid with two preceding ones and two following ones is a good enough beginning. The protein structure data accumulated in databases should help in such an analysis. Second, we need to guess the doublet code assignments from the known triplet code. This has already been studied to some extent within the context of the wobble rules, but it should be investigated in more detail keeping the constraints of the synthetase classes in mind.

VII. SUMMARY AND OUTLOOK

I have looked at the structure of proteins from an information theory point of view. The emphasis is on the 3-dimensional structure of the end-product, i.e. how should the segments of a polypeptide chain be chosen so that it folds into the required shape. The means used to achieve that end are secondary, i.e. which amino acids should be chosen so that the interactions amongst their R-groups make the polypeptide chain fold in the required manner. This emphasis is in sharp contrast with the conventional approach to the protein folding problem, i.e. find the 3-dimensional structure of the protein, given the sequence of amino acids and the interactions of their R-groups. The conventional problem is believed to be NP-hard, because finding the global energy minimum with all possible interactions is not at all easy. The rephrased problem of structural design may not be that hard—the local orientation of a building block can be fixed by its interactions with its neighbours; it is enough to have a locally stable or metastable configuration and not necessarily a global energy minimum [19]. Also, there are many ways a folded chain can cover a 3-dimensional shape, and quite likely there is a lot of flexibility in choosing the sequence of amino acids without substantially altering the structure of the protein.

The fundamental operations needed for processing structural information are translation and rotation. I have shown that carbon and its tetrahedral geometry provide the simplest discretisation of these operations. For the construction of proteins as folded chains, the polypeptide chain is the simplest back-bone containing rigid segments alternating with flexible joints. To fold this back-bone into arbitrary shapes on a diamond lattice requires 9 basic operations. The amino acids somehow implement these operations by interactions amongst their side chain R-groups.

I have pointed out that the division of the 20 amino acids, by aminoacyl-tRNA synthetases, into two classes of 10 each has structural significance. Every R-group property is equally divided between the two classes, such that the shorter side chains are in class II and the longer ones in class I. Combining this fact with the number of discrete operations required to fold a polypeptide chain, and the result that two quantum queries can distinguish 10 items, I have proposed that the present triplet genetic code was preceded by a primitive doublet one. How the doublet code was converted to a triplet one is a matter of conjecture, and I have outlined one possible scenario.

Knowing the solution selected by evolution has no doubt guided my logic. Still unravelling the optimisation criteria involved in the design of molecules of life is a thrilling exercise. It should be kept in mind that evolution has discovered its optimal parameters, not by logical deduction, but by trial and error experiments (of course using the available means). For that reason the chosen parameters are not always perfect. On the other hand, evolution has had plenty of time for experimentation, something which we do not and cannot have. As a result, though evolution is not perfect in finding its criteria, it is impressive to say the least!

In closing, I want to contrast different information processing paradigms. Electronic computation uses physical building blocks and operations based on real variables. Quantum computation extends the building blocks and operations to the complex numbers. Structural information processing goes still one step further, to the non-commutative algebra of quaternions. Systematic analysis of structural information processing has a long way to go. Yet in a sense, it came first—proteins arose before genes, nervous signals, spoken and written languages, number

systems and computers [20].

ACKNOWLEDGEMENTS

I am grateful to C. Ramakrishnan and M.A. Viswamitra for useful discussions regarding protein structure.

---

[1] A. Patel, `quant-ph/0012149`.

[2] The group of 3-dimensional rotations is $SU(2)$, which can be represented using matrices or quaternions.

[3] L. Grover, Proceedings of the 28th Annual ACM Symposium on Theory of Computing, Philadelphia (1996), p.212, `quant-ph/9605043`.

[4] G.N. Ramachandran, C. Ramakrishnan and V. Sasisekharan, J. Mol. Biol. 7 (1963) 95.

[5] A.L. Lehninger, D.L. Nelson and M.M. Cox, *Principles of Biochemistry*, Second Edition, Worth Publishers, USA (1993).

[6] T.E. Creighton (Ed.), *Protein Folding*, W.H. Freeman, New York (1992).

[7] Instead of an R-group, glycine has two hydrogen atoms attached to the C_α atom. That makes glycine achiral, but the number of attachment possibilities for the C_α atom remains one.

[8] The peptide bond is a little shorter than the single bonds and its bond angles of 120° are somewhat wider than the tetrahedral angle. These two deviations tend to compensate for each other.

[9] Carbon can also form a hexagonal diamond lattice, with the same tetravalent bonds and density as the face-centred cubic diamond lattice. Such hexagonal diamond crystals do not occur terrestrially, but they have been found in meteorites and have been synthesised in laboratory.

[10] G. Eriani, M. Delarue, O. Poch, J. Gangloff and D. Moras, Nature 347 (1990) 203.

[11] J.G. Arnez and D. Moras, Trends Biochem. Sci. 22 (1997) 211.

[12] B. Lewin, *Genes VII*, Oxford University Press (2000).

[13] A. Patel, Proceedings of the Winter Institute on Foundations of Quantum Theory and Quantum Optics, Calcutta (2000), Pramāṇa 56 (2001) 365, `quant-ph/0002037`.

[14] The non-integer number of items means that the algorithm has an intrinsic error. In the case of two queries and 10 items, the error rate is about 1 part in 1000.

[15] F.H.C. Crick, J. Mol. Biol. 19 (1966) 548.

[16] Symbiosis is not at all uncommon during evolution. There is evidence that the cellular organelles mitochondria and chloroplasts, with their own genetic material, first developed independently and were later incorporated in ancestral cells with eukaryotic nuclei.

[17] F.H.C. Crick, J. Mol. Biol. 38 (1968) 367.

[18] Another possibility, for instance, is that the transition from doublet to triplet code replaced three amino acids with short side chains cramped together by two amino acids with longer chains and expanded structure.

[19] Diamond is structurally the strongest material, but it is energetically metastable.

[20] The word "protein" derives from the Greek "protos" meaning "first" or "foremost".